

基于 DoG 掩模的冷冻电镜生物大分子图像特征提取

巫小蓉^{1,2} 吴效明³

(华南理工大学计算机科学与工程学院 广州 510641)¹

(广东外语外贸大学信息科学技术学院 广州 510420)² (华南理工大学生物工程学院 广州 510641)³

摘 要 针对冷冻电镜生物大分子图像低信噪比(SNR)和低对比度的特点,提出了基于高斯差分(DoG)掩模的形状特征提取方法。该方法利用高斯差分能提取多尺度梯度信号的优势,提取了目标生物大分子颗粒的近似区域,并在此基础上定义了由 14 个形状统计特征组成的特征向量。实验结果表明,该方法能有效提取生物大分子颗粒的形状特征,为进一步进行颗粒识别奠定了良好的基础。

关键词 颗粒识别,大分子三维重构,高斯差分,形状特征,特征提取

中图分类号 TP391 文献标识码 A

Feature Extraction from Cryo-EM Image Based on DoG

WU Xiao-rong^{1,2} WU Xiao-ming³

(School of Computer Science & Engineering, South China University of Technology, Guangzhou 510641, China)¹

(School of Information Science and Technology, Guangdong University of Foreign Study, Guangzhou 510420, China)²

(School of Biological Science & Engineering, South China University of Technology, Guangzhou 510641, China)³

Abstract It may be difficult to extract distinctive features pertinent to a specimen when dealing with very low-contrast and low SNR cryo-Electron micrograph(Cryo-EM) images. A DoG(difference of Gaussian) based method to extract shape features was provided in this paper, which is aimed to get a binary image by using DoG. This binary image was used as an object mask and shape features were calculated for the largest object in the mask. The experiments showed this method has good results.

Keywords Particle picking, Macromolecular 3D reconstruction, DoG, Shape feature, Feature extraction

人类基因组工作草图的绘制完成,标志着生命科学进入了后基因组时代。在人类基因组测定之后,人们发现仅从基因组序列的角度根本无法完整、系统地阐明生物体的功能,必须对蛋白质进行结构和功能上的研究,才能更好地理解生命过程。而以蛋白质为主体的生物大分子的功能主要决定于它们的三维结构。因此,研究生物大分子三维结构和功能之间的关系成为了现代生命科学的一个重要研究领域,生物大分子高分辨率三维结构的研究甚至原子水平结构的研究也已经成为现代结构生物学的主要内容之一。

冷冻电镜生物大分子三维重构方法采用冷冻电子显微技术(简称冷冻电镜技术, Cryo-EM)和计算机三维重构技术相结合的方式,是用来测定生物大分子三维结构的主要方法之一。该方法的主要思路是利用冷冻电镜和低剂量电子束成像技术对样本进行电子成像,利用高灵敏底片进行成像记录,并对记录图像进行图像处理,包括生物分子颗粒识别、校正、去噪等,最后完成样本的三维重构计算。

1 研究现状

从冷冻电镜图像中识别生物大分子颗粒是整个计算机三

维重构技术非常重要的环节,直接影响三维重构的效率和精度。生物大分子三维重构所需要的颗粒数量大,为了达到“原子”分辨率,需要大约一百万个颗粒投影。

目前,已有许多生物大分子颗粒自动识别方法,例如模板匹配方法^[1]、基于边缘检测和图像分割方法^[2]、基于神经网络方法^[3]等。但冷冻电镜图像由于质量差以及生物大分子颗粒的多样性,没有一种能被广泛使用。

基于特征识别的方法也是使用得最多的方法之一^[4-6]。该方法通常依赖于生物大分子颗粒图像的显著特征,如几何特征、统计特征等。考虑到生物分子颗粒具有不同旋转角度,特征必须具备旋转不变性。从低对比度的冷冻电镜图像中正确提取具有旋转不变性的特征是该方法所面临的难点。Sorzano^[4]将图像转换为极坐标图像,并依据环定义了具有旋转不变性的特征向量。Satya^[5]采用 5 类具有多尺度的基于积分图像的模板作为特征进行识别。Richard^[6]使用了 8 个基于灰度的统计特性,还提出了由方差图像的二进制图像作为掩模后计算得到的 8 个特征:目标对象的灰度均值和方差、大小和中心坐标、沿最长轴及与它成 90 度的轴的长度和周长。

到稿日期:2009-10-26 返修日期:2010-01-11 本文受国家自然科学基金项目(30670538),广东省高等院校学科建设专项资金资助。

巫小蓉(1977-),女,博士生,讲师,主要研究方向为模式识别、图像处理、生物信息,E-mail:wxr@mail.gdufs.edu.cn;吴效明(1950-),男,博士,教授,主要研究方向为生物医学信息、医学仪器、医学图像处理、生物力学。

但通过研究发现,在用方差图像得到的二进制掩模对颗粒内部具有明显纹理的图像并不理想,因而所提取到的特征信息的识别能力也就受到了限制。

本文对此情况,在 Richard 研究基础上提出了基于 DoG 掩模冷冻电镜生物大分子图像特征提取方法。该方法首先计算图像的高斯差分,根据阈值进行二进制转化,并以此作为掩模与原图像进行与操作得到图像颗粒近似区域。此外,本文还定义了更具识别能力的形状特征向量。

2 基于 DoG 掩模的特征提取

2.1 冷冻电镜图像的特点

冷冻电镜获得的是处于天然状态下未经染色的生物分子的二维投影像。由于样本未经染色,很容易受到电子束的损伤。将样本快速冷冻(样本处于玻璃态)并在低温下观察样本,一定程度上可以减少电子束对样本的损伤,因此称为冷冻电镜。为了使样本能够在整个成像过程中受到的辐射总剂量损伤最小,在收集电镜显微图像时,电子束被局限于很小的剂量。这样小的剂量导致图像质量很差。冷冻电镜图像具有低信噪比、低对比度、背景强度不均衡、颗粒内部纹理不规则等特点。

2.2 基于 DoG 算子的掩模提取

在冷冻电镜图像中为了识别生物分子颗粒,需要屏蔽颗粒内部或背景中的细小信号。多尺度空间方法模拟了人类在不同距离观察事物的过程。通常大尺度对应图像的概貌,而小尺度对应图像的细节特征。Koenderink 已经证明高斯卷积核是实现尺度变换的唯一变换核^[7]。二维高斯函数可以定义为

$$G(x, y, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x^2+y^2)}{2\sigma_k^2}} \quad (1)$$

式中,标准差 σ_k 是高斯滤波器的唯一参数,也称为尺度因子。它与滤波器操作邻域的大小成正比。 (x, y) 是空间坐标。高斯平滑的实际含义是可以可靠地发现边缘。如果只需要全局性的显著边缘,可以增大高斯平滑滤波器的尺度因子 σ_k ,使得比较不明显的特征得以抑制。

本方法将先从整幅冷冻电镜图像中提取候选颗粒图像,包括含有颗粒的和不含颗粒的。将候选颗粒图像表示为 $I(x, y)$,则其不同尺度下的高斯变换如式(2)所示。

$$L(x, y, \sigma_k) = G(x, y, \sigma_k) * I(x, y) \quad (2)$$

式中, (x, y) 表示图像的像素位置, σ_k 是高斯函数尺度因子。DoG 是两个不同尺度的高斯变换的差值,如式(3)所示。

$$D(x, y, \sigma) = L(x, y, \sigma_i) - L(x, y, \sigma_j) \quad (3)$$

式中, σ_i, σ_j 是两个不同的尺度因子。将图像 $D(x, y, \sigma)$ 通过一定的阈值,变换为二进制图像。将该二进制图像作为掩模与原图像进行与操作,然后提取最大的独立区域作为近似对象。图像特征从该近似对象提取。

2.3 特征提取

为了识别冷冻电镜图像中不同旋转角度的生物大分子颗粒,提取的特征必须具有旋转不变性。本文定义了 14 个具有旋转不变性的特征。首先是 7 个 Hu 矩和扁度, Hu 矩满足平移、旋转不变性,被广泛应用于区域形状识别中,而扁度可以

近似模拟区域的长短轴之比。其次是近似对象的方差,用以衡量近似对象的内部纹理。另外 5 个包括近似对象中心偏移,面积 S , 周长 L 以及圆形度 R 和形状复杂性 e 。中心偏移定义为重心与候选颗粒图像中心坐标距离,衡量对象是否位于图像中心。圆形度 R 定义如式(4)所示, R 值越大,则区域越接近圆形,该特征对区分圆形对象很重要。

$$R = 4\pi S / L^2 \quad (4)$$

形状复杂性 e 定义如式(5)所示, e 值越大,则表明单位面积的周长越大,区域离散,对象形状复杂性高,反之,则为简单形状。

$$e = L^2 / S \quad (5)$$

3 实验分析

3.1 实验数据和方法

本文采用 keyhole limpet hemocyanin (KLH) 病毒颗粒冷冻电镜图像作为实验数据。该数据集是国际上作为生物大分子颗粒自动识别的公共基本测试数据,可从如下网站获得 http://ami.scripps.edu/prtl_data/klh/klh_1k/。

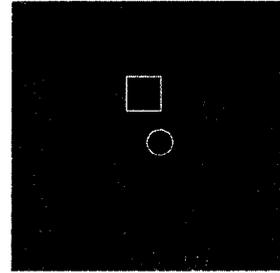


图1 KLH 病毒冷冻电镜图像

图 1 是数据集中的一幅,原图像大小为 $2048 * 2048$ 像素。图 1 中的白色方框标注的方形颗粒是 KLH 的侧面投影,也是要识别的颗粒投影;而白色圆圈标注的圆形颗粒是 KLH 的顶部投影,这些圆形颗粒以及其他一些不规则或堆叠的均是不需要的。

对冷冻电镜图像先进行去噪等预处理,然后采用交叉相关方法将待识别的候选颗粒图像提取出来,提取出来的候选颗粒图像大小为 $256 * 256$ 像素。再对候选颗粒用本文方法进行特征提取。为了说明本文方法的优势,将用差分图像掩模方法^[6]进行对比分析。

3.2 实验结果及分析

本文选取 60 个候选颗粒作为研究对象,分别用基于 DoG 掩模方法和基于差分图像掩模方法进行处理。由于篇幅所限,仅列出 3 个候选颗粒进行比较,其中第一个颗粒是正确颗粒。图 2 是本文方法处理结果,其中第一列是候选颗粒原图,第二列是高斯差分图像,第三列是第二列进行二进制化后的掩模,第四列是提取的第三列中最大独立对象并经过填充内部空洞后的掩模,第五列是最后得到的近似颗粒,这也是进行特征提取的图像。图 3 是 Richard 方法处理结果,与图 2 的不同之处是图 3 的第二列是原图的差分图像平滑后得到的效果。从两种方法的处理效果来看, DoG 方法能更好地区分不同尺度的梯度信息,得到的近似颗粒更接近实际颗粒形状,因而由此提取的特征值也将更准确。

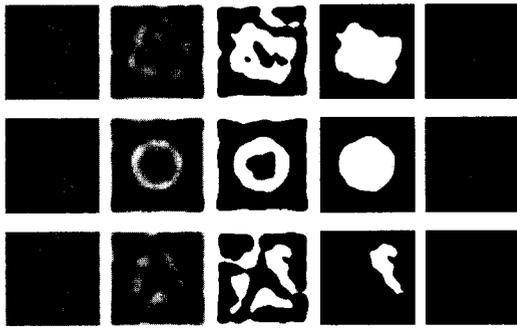


图2 3个候选颗粒图像采用DoG掩模的处理结果

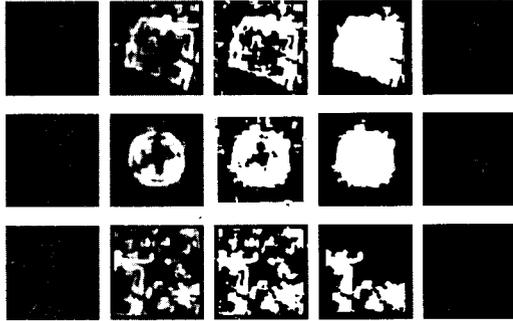


图3 3个候选颗粒图像采用差分图像掩模的处理结果

表1 DoG和差分图像掩模特征识别率对比

特征	DoG	差分图像
面积	68%	36.96%
中心偏移	85%	44.74%
形状复杂性	41.46%	36.96%
周长	43.59%	45.95%
Hu矩M1	70.83%	33.33%
Hu矩M2	51.52%	30.91%
Hu矩M3	56.67%	29.82%
Hu矩M4	44.74%	32.69%
Hu矩M5	43.59%	31.48%
Hu矩M6	45.95%	33.33%
Hu矩M7	47.22%	31.48%
扁度	53.13%	33.33%
方差	50%	43.59%
圆形度	41.46%	36.96%

表1列出了本文定义的14个特征在两种方法下的单一识别率。特征的单一识别率是仅依据该项特征进行分类正确的百分比。从结果可以看出,本文提出的方法计算得到的特

征值的单一识别率是令人满意的。除了周长以外,本文方法得到的特征值的识别能力均高于差分图像掩模方法。在此基础上,如能选用合适的分类器,将取得良好的识别效果。

结束语 特征是模式识别的关键。形状特征是识别生物大分子颗粒的有力依据。冷冻电镜图像因为低对比度和低信噪比的特点,很难提取其形状特征。DoG能提取不同尺度的边缘信号。本文利用DoG这一特点提取生物大分子颗粒的边缘信号,尽量抑制颗粒内部不明显的纹理细节信息,从而得到生物大分子颗粒的近似区域,并以此计算其近似形状特征。实验表明该方法得到的形状特征值对识别大分子颗粒能起到很好的效果。但本文方法中得到的近似颗粒区域还是会受到相邻物体的边缘影响而产生毛刺,因此,在今后的研究中还应逐步改进,以得到更接近实际颗粒的近似颗粒区域。

参考文献

- [1] Chen J Z, Grigorieff N. SIGNATURE: a single-particle selection system for molecular electron microscopy[J]. J. Struct. Biol., 2007, 157(1): 168-173
- [2] David W, Ben H, Ericksson G. The Laplacian of Gaussian and arbitrary z-crossings approach applied to automated single particle reconstruction[J]. Journal of Structural Biology, 2007, 159(1): 122-134
- [3] Ogura T, Sato C. Automatic particle pickup method using a neural network has high accuracy by applying an initial weight derived from eigenimages: a new reference free method for single-particle analysis[J]. J. Structure Biology, 2004, 145: 63-75
- [4] Sorzano C O S, Recarte E, Alcorlo M, et al. Automatic particle selection from electron micrographs using machine learning techniques[J]. Journal of Structural Biology, 2009
- [5] Satya P M, Zhu Yuanxin, Kriegman D. Detecting particles in cryo-EM micrographs using learned features[J]. J. structure Biology, 2004, 145: 52-62
- [6] Hall R J, Ardan P. A two step approach for semi-automated particle selection from low contrast cryo-electron micrographs[J]. J. Structure Biology, 2004, 145: 19-28
- [7] Pizer S M, Koenderink J J, Lifshitz L M, et al. An image description for object definition based on external regions in the stack [C]//Proc. of 9th Conf. Information Processing in Medical Imaging. Boston, MA: Martinus Nijhoff, 1986

(上接第269页)

- [2] Elad M, Aharon M. Image denoising via sparse and redundant representations over learned dictionaries[J]. IEEE Transactions on Image Processing, 2006, 15(12): 3736-3745
- [3] Protter M, Elad M. Image Sequence Denoising via Sparse and Redundant Representations[J]. IEEE Transactions on Image Processing, 2008, 18: 27-35
- [4] Bryt O, Elad M. Compression of Facial Images Using the K-SVD Algorithm[J]. Journal of Visual Communication and Image Representation, 2008, 19(4): 270-283
- [5] Liao H, Sapiro G. Sparse image representation for limited data tomography[J]. IEEE International Symposium on Biomedical Imaging, 2008: 1375-1378
- [6] Donoho D, Elad M. Optimal Sparse Representation in General Dictionaries via L1 Minimization[J]. Proceedings of the National Academy of Sciences, 2003, 100: 2197-2202
- [7] Blumensath T, Davies M. Gradient pursuits [J]. IEEE Transac-
- tions on Signal Processing, 2008, 56(6): 2370-2382
- [8] Cande E. Compressive Sampling [C]//Proc. Int'l Congress of Mathematicians, 2006
- [9] Marcellin M W, Gormish M J, Bilgin A, et al. An overview of JPEG-2000[C]//Proc. Data Compression Conf. 2000: 523-541
- [10] Starck J L, Candes E J, Donoho D L. The curvelet transform for image denoising[J]. IEEE Trans. Image Process., 2002, 11: 670-684
- [11] Gstaad R, Starck J L. Dynamic range compression: A new method based on wavelet transform[C]//Astron. Data Anal. Software Systems Conf. Strasbourg, 2003
- [12] Starck J L, Elad M, Donoho D L. Image decomposition: Separation of texture from piece-wise smooth content[C]//SPIE Conf. Signal Image Process. Wavelet Applicat. Signal Image Process. 2003
- [13] Wright J, Yang Y, Ganesh A, et al. Robust face recognition via sparse representation[J]. pattern analysis and machine intelligence, 2009, 31(2): 210-227